The relationship between response consistency in picture naming and storage impairment in people with semantic variant Primary Progressive Aphasia

Cornelia van Scherpenberg[1,2 *]

Nora Fieder[1,3]

Sharon Savage[3,4]

Lyndsey Nickels[3]

[1]Berlin School of Mind and Brain, Humboldt Universität zu Berlin, Germany

[2]Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

[3]ARC Centre of Excellence in Cognition and its Disorders, Department of Cognitive Science, Macquarie University, Sydney, Australia

[4]Psychology Department, University of Exeter, UK

* Corresponding author at Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, Luisenstraße 56, 10117 Berlin, Germany, e-mail: cornelia.vanscherpenberg@hu-berlin.de

van Scherpenberg, C., Fieder, N., Savage, S., & Nickels, L. (in press). The relationship between response consistency in picture naming and storage impairment in people with semantic variant Primary Progressive Aphasia. *Neuropsychology.*

Abstract

Objective. The progressive loss of stored knowledge about word meanings in semantic variant Primary Progressive Aphasia (svPPA) has been attributed to an amodal "storage" deficit of the semantic system. Performance consistency has been proposed to be a key characteristic of storage deficits but has not been examined in close detail and larger participant cohorts. ***Methods:*** We assessed whether 10 people with svPPA showed consistency in picture naming across three closely consecutive sessions. We examined item-by-item consistency of naming accuracy and specific error types, while controlling for the effects of variables such as word frequency, familiarity and age of acquisition. ***Results:*** Participants were very consistent in their accurate and inaccurate responses over and above any effects of the word-related variables. Analyses of error types that compared consistency of semantic errors, correct responses and other error types (e.g., phonologically related errors, unrelated errors) revealed lower consistency. ***Conclusions:*** Our findings support the assumption that semantic features constituting semantic representations of objects are progressively lost in people with svPPA and are therefore consistently unavailable during naming. Variability in the production of error types remains when distinctive features of an object are lost resulting in the selection of semantically or visually similar items, or in the failure to select an item and the production of a no-response. The assessment of performance consistency sheds light on the underlying impairment of people with semantic deficits (semantic storage versus access deficit). This can support the choice of an appropriate treatment technique aiming to maintain, or re-learn semantic information.

Keywords: semantic variant Primary Progressive Aphasia, storage impairment, consistency, naming, semantic features

Introduction

Semantic variant Primary Progressive Aphasia (svPPA; Gorno-Tempini et al., 2011; Hodges, Martinos, Woollams, Patterson, & Adlam, 2008; Hodges, Patterson, Oxbury, & Funnell, 1992) is a neurodegenerative disease which primarily affects language production and comprehension. Core diagnostic features include impaired confrontation naming and impaired single-word comprehension (Gorno-Tempini et al. (2011). Other cognitive skills outside of language, including working memory, episodic memory, orientation, problem solving and visuospatial skills remain relatively preserved until severe stages of the disease (Patterson & Hodges, 2000). On neuroimaging, individuals with svPPA show atrophy in the anterior temporal lobe which is usually more dominant left-laterally at first, but in later stages becomes bilateral (Acosta-Cabronero et al., 2011; Brambati et al., 2009; Gorno-Tempini et al., 2011; Mummery et al., 2000).

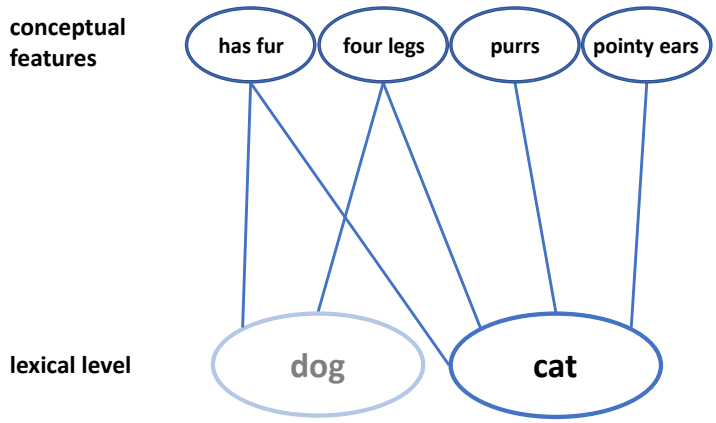
Storage deficits and response consistency in svPPA

There is now a substantial body of evidence supporting the theory that the language symptoms of svPPA are due to a central (amodal) impairment of the semantic system (e.g., Bozeat, Lambon Ralph, Patterson, Garrard, & Hodges, 2000; Coccia, Bartolini, Luzzi, Provinciali, & Lambon Ralph, 2004; Hodges, Graham, & Patterson, 1995; Marques & Charnallet, 2013; Patterson & Hodges, 2000). Moreover, this impairment in svPPA has generally been described as a "storage deficit" where stored semantic information is progressively lost (Hodges et al., 1995; Mirman & Britt, 2014; Shallice, 1987).

Feature-based theories of semantic representations assume that this semantic information is stored in the form of functional and perceptual features or attributes, which characterise word meanings and distinguish them from each other (e.g., Dell, Schwartz, Martin, Saffran, & Gagnon, 1997; Vigliocco, Vinson, Lewis, & Garrett, 2004). In tasks such as picture naming, these distinguishing features necessarily have to be activated to fully and

correctly retrieve the lexical representation (e.g., cat) and therefore name the picture "cat" (Figure 1a).

a



b

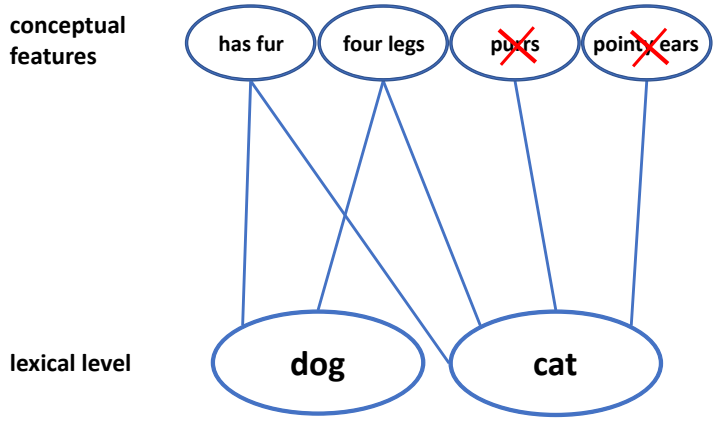


Figure 1. Schematic account of the semantic representation of "cat". Panel 1a: The unimpaired language system; Panel 1b: Loss of distinctive features which results in a coordinate error.

In svPPA, it is assumed that certain semantic features deteriorate and eventually are lost, and therefore lexical-semantic representations cannot be retrieved successfully (Hodges et al., 1995; Laisney et al., 2011). According to this assumption, participants with svPPA should experience a consistent failure to retrieve the correct lexical representation of a target in tasks such as picture naming.

If features are lost that are necessary to distinguish between concepts, for example, cat from dog, the participant might name a picture of a cat as "dog" instead (Figure 1b). This is usually referred to as a coordinate semantic error – an error type which has been commonly reported in svPPA (e.g., Budd et al., 2010; Jefferies & Lambon Ralph, 2006). If only those features are left that can identify a cat as an animal, then the picture might be named with the superordinate semantic error "animal". This response type is another frequent occurrence in picture naming in svPPA and becomes more frequent the further the impairment progresses (Budd et al., 2010; Hodges et al., 1995).

However, word finding difficulties and/or difficulties in word comprehension can also occur as a result of an impairment in the "access" to representations in the semantic system and/or in the access to/from semantic representations to word forms in the lexicon as often seen in people with stroke aphasia (e.g., Warrington & Shallice, 1979). In this case, semantic representations themselves are intact but cannot be accessed reliably. This may occur in the early stages of svPPA, where comprehension is still relatively intact but individuals appear to have reduced naming abilities due to weakened access to semantic representations in language production (Mesulam et al., 2009; Wilson, Dehollain, Ferrieux, Christensen, & Teichmann, 2017).

Warrington and Shallice (1979) proposed that one of the main characteristics that can be used to distinguish between a storage deficit and an access deficit is response consistency. They suggested that if a representation is 'lost' due to a storage deficit it should be consistently unavailable – in the same task over time and in tasks assessing different verbal and non-verbal modalities. In contrast, if representations are still present but inaccessible due to an access deficit, retrieval of those representations might be less consistent due to fluctuations in their accessibility. Evidence for consistency in participants with a putative storage deficit comes from Chertkow and Bub (1990) who found high consistency of accurate and inaccurate responses in two sessions of a picture naming task with 10 participants with

Alzheimer's disease. Similarly, Warrington and Cipolotti (1996) reported several word-picture matching experiments with four participants with fronto-temporal dementia, who were given a clinical diagnosis at the time of probable Pick's disease, but fulfil the criteria for svPPA. This study showed that the responses of these participants to each item across the experiments was highly consistent (see also Coughlan & Warrington, 1981). Our study seeks to confirm and extend these findings: Given that individuals with svPPA are claimed to suffer from a storage deficit with semantic information being lost, this predicts a permanent and consistent inability to name specific words correctly (cf. Figure 1).

Evidence for performance inconsistency due to an access deficit was presented in a study by McCarthy and Kartsounis (2000) which described participant FAS who suffered from an impairment at a post-semantic but pre-phonological processing level. Across four testing sessions, a third of all the test items that were named by FAS were highly variable in their accuracy: FAS could name these items correctly on one but not on another occasion.

Critically, when evaluating consistency, it is vital that other factors are also controlled. For example, Howard (1995) demonstrated that a considerable amount of the consistency shown in lexical retrieval by participant EE was due to word familiarity, with highly familiar words being consistently named correctly. Hence, other psycholinguistic variables that can influence performance, such as familiarity or frequency, that can lead to some items being consistently named accurately or inaccurately need to be controlled (see also Rapp & Caramazza, 1993).

Further evidence for a distinction in consistency between storage and access deficits comes from a study by Gotts and Plaut (2002) who used computational modelling to simulate the symptoms of access and storage deficits. They simulated a storage deficit as damage to connections between neuron-like semantic units (Buszáki, 2010; Hebb, 1949; McClelland & Rogers, 2003). The higher the proportion of connections that were lesioned (i.e., removed) the more information was lost. An access deficit was simulated by increasing the level of

activation needed to activate and access the units that code semantic information. The model was only able to generate response consistency when connections were severely damaged in case of a storage deficit. This suggests that with different disease severity we may expect different patterns of consistency.

Previous studies investigating response consistency in people with svPPA have mostly examined consistency across modalities: For example, Bozeat et al. (2000) and Jefferies and Lambon Ralph (2006) reported high item-by-item consistency across picture naming, word-picture matching and sound-matching. They also found sensitivity to concept familiarity in their participants, with familiarity emerging as a significant predictor of the degree of consistency shown. The only study that has examined item consistency in the same task over time was conducted by Hodges et al. (1995) who reported high item-by-item consistency on accuracy for JL, a man with svPPA. JL named almost all pictures consistently incorrectly over five sessions with five to six months between sessions. Unfortunately, there was no examination of the extent to which familiarity and frequency could have accounted for this consistency, nor whether there was also consistency between error types within the incorrect responses. Thus, while these studies provide evidence for a storage deficit in the semantic system in svPPA, there remains a need to study consistency of performance over time in a larger sample of participants while controlling for familiarity, frequency and other semantically and lexically relevant variables. Moreover, while items might be consistently named incorrectly, it is important to also consider error type consistency.

In the present study, we address these issues by assessing item-by-item consistency within modality over repeated presentations in a case series of ten people with svPPA. Our sample size matches or even exceeds that of most studies which have previously studied consistency and storage impairments (e.g., Bozeat et al., 2000 (n = 10); Coughlan & Warrington, 1981 (n = 1); Hodges et al., 1995 (n = 1)) or other aspects of language production

in svPPA (e.g., Marques & Charnallet, 2013 (n = 6); Montembeault et al., 2017 (n = 9)). We not only examined consistency between correct/incorrect responses but also between semantic errors and other error responses, while at the same time controlling for effects of psycholinguistic variables such as familiarity and frequency on naming. By doing so we aimed to establish whether consistency can indeed be used as a criterion to distinguish a "storage" deficit from an "access" deficit even when including a finer distinction within incorrect responses. Moreover, the characteristics of our participant cohort allows us to investigate consistency at different severities of svPPA.

Methods

Participants

The current study included 10 participants with svPPA who took part in a word-relearning study by Savage and colleagues (Savage, Ballard, Piguet, & Hodges, 2013; Savage, Piguet, & Hodges, 2014, 2015). All participants were originally recruited through FRONTIER, the Frontotemporal Dementia Research Group clinic at Neuroscience Research Australia, Sydney, and had been diagnosed with svPPA according to the consensus criteria (Gorno-Tempini et al., 2011) by an experienced behavioural neurologist (Hodges), based upon detailed clinical assessment, neuropsychological assessment and, where possible, structural brain magnetic resonance imaging (see Table 1). Participant labels have been retained from Savage et al. (2015), where participants B1, C2, G1, J1, S1 and T4 were reported; J2 and J3 in this study are SD-J2 and SD-J3 in Savage et al. (2014) and K1 was participant SD3 in Savage, Ballard, et al. (2013). We selected all participants for whom data was available from three pre-therapy naming baselines, with a minimum of 50% overt responses in each baseline, to allow for statistical analysis. Demographic information is summarised in Table 1.

Table 1**:** *Participant demographic information.*

Participants	**J3**	**T4**	**C2**	**G2**	**K1**	**J2**	**G1**	**B1**	**S1**	**J1**
Sex	M	M	M	F	M	F	M	M	F	M
Age (yrs)	56.2	63.9	50.3	57.8	66.8	71.4	63.3	61.9	62.3	69.5
Years of Education	11	11.5	12	15	9	15.5	16	13	14.5	15
First Language	Portu-guese*	Engl.	Engl.	Engl.	Engl.	Engl.	Engl.	Engl.	Engl.	Engl.
Handedness	right	left	right	right	right	right	right	right	right	right
Disease Duration (yrs)	7.8	6.7	8.3	4.5	7.3	9.2	6.8	5.2	6.3	5.4
Focus of Temporal lobe atrophy	left	left	right	left	NA	left	left	left	left	left

*J3 learned English when he was 9 years old and, lived in an English-speaking country since that time and had 3 years of formal education in English from age 14-17.
M: Male; F: Female; Engl.: English. NA: MRI not available due to presence of a pacemaker

Magnetic Resonance Images obtained for 9 participants (excluding K1 who had a pacemaker and therefore could not be scanned) showed the typical pattern of atrophy in the anterior temporal lobe. All but one of the participants showed predominantly left temporal lobe atrophy. Although participant C2 had dominant right temporal lobe atrophy, he did not perform differently on any of the tasks investigated (see below). For details of imaging and atrophy for all participants see the original papers (Savage, Ballard, et al., 2013; Savage et al., 2014, 2015).

Table 2**:** *General cognitive and language related abilities, ordered by severity.*

Participants	**J3**	**T4**	**C2**	**G2**	**K1**	**J2**	**G1**	**B1**	**S1**	**J1**	**cut-off scores**
General Cognitive Assessments											
ACE-R[a] Total (100)	45*	49*	57*	62*	68*	56*	68*	84*	80*	86	82
ACE-R[b] Subtotal Language (26)	9*	6*	11*	11*	14*	11*	14*	16*	18*	21*	22.1
ACE-R[c] Category Fluency: Animals	3*	3*	10*	5*	12*	13*	7*	12*	10*	15*	NA
Digit Span[d] Total (Age scaled score)	13	7	8	10	12	8	10	15	7	18	
Forward (max. span)	7	5	8	6	7	5	8	9	5	8	4.6
Backward (max. span)	5	4	5	4	5	4	4	6	4	8	2.9
Rey Complex Figure[e] Copy Score (36)	36	34	30	35	34	35	36	35	28*	34	28.2
Three-Min-Delay Score (36)	27.5	12	9	18.5	2*	25.5	22.5	17	19.5	23	6.9
Trail-Making-Test[f] Trail A (secs)	33	31	39	55	42	35	34	39	39	29	NA
Trail B (secs)	113	88	78	81	64	111	80	41	73	61	NA
Language Assessments											
SYDBAT[g] Naming (30)	1*	2*	4*	6*	6*	8*	8*	10*	16*	16*	22
Repetition (30)	25*	25*	29	30	30	28*	29	30	30	29	29
Comprehension (30)	13*	12*	14*	20*	22*	17*	17*	19*	26	27	26
Semantic (30)	12*	7*	15*	13*	21*	15*	17*	19*	26	25	24
TROG[h] Total Score (80)	68	62	71	71	76	79	74	78	76	80	NA
TROG Block Score (20)	12*	9*	12*	14*	17	19	16	18	18	20	14
FRS[i] **Rasch Score**	-1.84	-1.84	-0.8	-0.2	1.68	3.35	1.47	NA	1.68	0.16	
FRS[j] **stage**	severe	severe	severe	moderate	moderate	mild	moderate	NA	moderate	moderate	

Maximum scores in brackets. Cut-off scores, where available, are published scores from normative test data, or represent scores that would be considered the limit of 'normal' at two standard deviations below the control mean. NA: Not available

[a] Addenbrooke's Cognitive Examination Revised (ACE-R)(Mioshi, Dawson, Mitchell, Arnold, & Hodges, 2006) is a brief test battery for evaluating performance in 5 cognitive domains (attention, memory, fluency, language and visuospatial tasks). Controls had a mean score of 93.7, see Mioshi et al. (2006).

[b] The language subtest of the ACE-R assesses reading, writing, comprehension, repetition and semantic tests.

[c] Category fluency – participants are asked to name as many animals as possible within 60 seconds, testing their ability to produce members of a certain semantic category. Controls name 17 category members on average (Marczinski & Kertesz, 2006; Savage, Hsieh, et al., 2013).

[d] Digit Span subtest of the Wechsler Adult Intelligence Scale (Wechsler, 2008) – assesses verbal working memory. Participants have to repeat a series of digits forward and backwards. Control data obtained by Savage, Hsieh, et al. (2013) yields an average forward digit span of 7.2 (SD = 1.3) and a backward digit span of 5.5 (SD = 1.3).

[e] Rey Complex Figure Test (RCFT; Meyers & Meyers, 1995) assesses visuospatial skills and short-term episodic memory. First, the participant is asked to copy the figure, and then to reproduce the same figure from memory after a 3-minute delay. Controls score a mean of 33.76 (SD = 2.8) on the copy task and 17.33 (SD = 5.2) on the delayed task, see Strauss, Sherman, and Spreen (2006, p. 828).

[f] Trail-Making-Test (see Strauss et al. (2006, pp. 655-677) – is a two-part test involving visual search, , speed of processing, and mental flexibility. In Part A, participants draw a line to connect numbers in sequential order. In Part B, they have to alternate between numbers and letters (e.g., 1 –A – 2 – B etc.). Scores are derived from the time needed for completion, which requires speeded performance and close attention. Tombaugh (2004) presents normative data stratified by age and education.

[g] Sydney Language Battery (SYDBAT) (Savage, Hsieh, et al., 2013) assesses language production and comprehension at a single word level to distinguish subtypes in Primary Progressive Aphasia. The SYDBAT comprises 4 subtests: 1) naming: participants name 30 pictures of objects with decreasing word frequency; 2) repetition: participants repeat single words after the examiner; 3) comprehension: participants match a spoken target word to one out of seven pictures; 4) semantic: participants select the closest match to a target picture from four semantically related pictures.

[h] Test for the Reception of Grammar (TROG) (Bishop, 1989) - participants match sentences with increasing syntactic complexity to one out of four pictures. The task has twenty blocks each with four samples of a particular syntactic construction, all four of a block must be answered correctly to 'pass' a block.

[i] Frontotemporal Dementia Rating Scale (FRS; Mioshi, Hsieh, Savage, Hornberger, & Hodges, 2010).

[j] Based on the FRS Rasch-converted score rating behavioural and cognitive abilities of people with frontotemporal dementia, 6 severity classes were identified: very mild, mild, moderate, severe, very severe, profound (Mioshi et al., 2010).

[*] where normative data was available, * indicates results outside the normal range

Cognitive screening and classification. All of the participants underwent extensive clinical and cognitive assessments. A detailed summary of each participant's cognitive and language abilities is shown in Table 2. As described in previous publications, all participants showed marked deficits in word finding (anomia), as seen on their spoken picture naming (range 1-16/30 correct;) and category fluency performance (range 3-15 category members). All participants, except for the two with the least impaired naming, J1 and S1, also showed impaired general semantic processing and impaired language comprehension (SYDBAT). Most participants were still able to comprehend grammatically complex structures relatively well (TROG). Hence, consistent with the diagnosis of svPPA, the results of the language assessments supported a central semantic deficit as the impairment underlying the difficulties in word finding and language comprehension while lexical, grammatical and post-lexical processes remained mostly intact.

Scores on tests of general cognitive abilities were reduced only in participants with more severe svPPA. All of the participants had an intact verbal working memory as shown in the results of the digit span task and showed intact short-term episodic memory and preserved visuospatial skills on the Rey Complex Figure. Psychomotor speed was not reduced with almost all participants' results on the Trail-Making-Test Part A falling within the time span needed by controls in the respective age groups. Mental flexibility also remained intact, with only one of the more severe participants performing slower than controls on this task (J3: 113).

Stimuli

Stimuli were those used to create individualised word retraining programmes, as reported in earlier studies by Savage and colleagues (Savage, Ballard, et al. (2013); Savage et al. (2014, 2015). As a result, items included were relevant to each participant's individual language abilities, interests and everyday needs from the following semantic categories: animals, food, kitchen items, bathroom items, household items, tools, and clothes. The

majority of words were typical members of their category (e.g., tomato, spoon, vacuum cleaner) with the exception of a few very specific items such as "circular knitting needles" or "African buffalo". Each participant was tested on approximately 100 words (range: 79-129; varying depending on the severity of the disease), using photographs obtained from stock images or taken by a family member in order to depict the specific object that was used at the participant's home. For further detail of the methods behind item selection, see Savage, Ballard, et al. (2013); Savage et al. (2014, 2015).

Across participants, there were a total of 296 different words, of which 267 words were used for analysis (see Appendix A): Items removed included those where the participant's response had not been recorded due to a fault in the program, or when the object was highly specific and not a single word (such as "African buffalo", see above). Psycholinguistic variables were collated for each word: spoken word frequency as measured by the SUBTLEX UK corpus (van Heuven, Mandera, Keuleers, & Brysbaert, 2014) and number of syllables, number of phonemes. In addition, ratings were obtained from 20 healthy participants (Macquarie University students) for concept familiarity, age of acquisition and imageability using instructions from Gilhooly and Hay (1977) for age of acquisition, Pavio, Yuille, and Madigan (1968) for imageability, and Alario and Ferrand (1999) for familiarity. However, the use of individual pictures for each participant made it infeasible to obtain measures of visual complexity and name agreement for each picture.

Procedure

Participants were tested individually in their homes. The data used in this study were from the series of baseline tests conducted to establish each participant's naming performance prior to training. Participants were instructed to name each picture with a single word. In each session, the participants were presented with their personal pictures in random order. Each picture remained on the screen for 10 seconds. Sessions were audio recorded and subsequently transcribed and scored for accuracy by the experimenter. This baseline

assessment was repeated several times per week for a minimum of 4 sessions in total (Savage et al., 2014).

For the current study, we chose 3 consecutive sessions of baseline testing for each participant. We excluded the first session to reduce the difference in task familiarity across sessions, and when possible, analysed the second, third and fourth recorded baselines or the earliest sessions without data recording errors. For three participants (J1, J3, S1) the analysed sessions were recorded on consecutive days, for the remaining participants there was at least one day in between each session. For participants K1 and G2, the second and third session that we chose for analysis were not consecutive in the actual baseline tests, as two sessions had to be excluded due to data errors.

Error Coding

The responses were coded by the first author as correct or erroneous using an unpublished coding system devised by Best, Nickels, and Williamson (2005). Half of the data (50%) was double coded by an additional expert coder. Interrater agreement was 91.5%, with any disagreements discussed and recoded by consensus.

Only the first full response by the participant was scored (i.e. false starts were not coded, e.g. for the response fee… feet, only 'feet' was coded, nor were part responses that cannot stand alone as an English syllable such as single consonants, short vowels or CV with a short vowel).

Responses were scored as correct when they were identical to the target word, or when they were acceptable part responses (e.g., "phone" for "telephone"), acceptable alternative names for the picture (e.g., "saucepan" for "pot") or contained the target but with additional information (e.g., "some kind of pasta" for "pasta", "a big cat" for "cat").

Incorrect responses were classified as shown in Table 3. Semantically related responses were subdivided into one of seven types of relationship with the target.

Phonologically related, unrelated and visual errors were grouped as 'other error response' for purposes of statistical analysis.

Table 3: *Error coding scheme (Best et al. (2005)*

Categories of error		Example
	A. Semantic errors	
1.	Semantic - superordinates	category names e.g. animal for dog
2.	Semantic - coordinates	Single word responses within the same category e.g. "glass" for "cup"
3.	Semantic - subordinates	e.g. "iphone" for "phone"
4.	Semantic - associatively related error responses	response and target share a semantic context but not semantic category e.g., "kitchen" for "stove"
5.	Semantic - semantic descriptions	Multi-word responses e.g., "gadgets for adding" for "calculator")
6.	Semantic - information from episodic memory	e.g., "that's the fruit that my father grew" for "cucumber"
7.	Semantic - semantic other	responses that shared a semantic relationship with the target but did not fall into any of the above categories e.g., false superordinates such as "vegetable" for "watermelon"
	B. No specific response	Responses such as "can't remember", "don't know", or omissions
	C. Other Responses	
A.	Other- Phonologically related	semantically unrelated and contained greater than or equal to 50% of the target word's phonemes regardless of whether they were words or nonwords
B.	Other - Unrelated responses	semantically unrelated words or nonwords containing less than 50% of the target word's phonemes, or vice versa
C.	Other - Visual errors	misperceptions of the depicted object (e.g., "spoon with ice" for a spoonful of sugar; or naming objects other than the target in the picture e.g., "table" for "placemats")

Analysis

Given the individualised nature of the item lists, each case was examined separately in individual analyses. A supplementary analysis was then run to assess whether patterns of consistency/inconsistency were present homogeneously for the group. The statistical analyses were performed using R version 3.3.1 (R Core Team, 2016).

Individual analyses

Consistency across sessions for each participant was assessed in two main analyses:

A. Cohen's Kappa (Cohen, 1960) to assess the consistency of accuracy of individual items (i.e. the degree to which the same items were consistently accurate or inaccurate across consecutive sessions), and to assess how consistently participants named an item using a certain error type across the three sessions. Since semantic errors were the most common overt error type across all participants, and are a key diagnostic symptom of svPPA, we compared the consistency of naming an item with a semantic error in Sessions 1, 2 and 3 compared to naming it accurately or with any other error type. Strength of agreement was interpreted using ranges provided by Landis and Koch (1977): Kappa: < 0.00 (poor), 0.00-0.20 (slight), 0.21-0.40 (fair), 0.41-0.60 (moderate), 0.61-0.80 (substantial), 0.81-1.00 (almost perfect).

B. Logistic regressions and multinomial regressions were used to determine whether the naming responses of Session 3 could be better predicted when responses in either Session 1 or 2 were added as an independent variable, over and above the predictive value that other psycholinguistic variables (familiarity, age of acquisition, imageability, frequency and number of phonemes) may have on naming performance (Hirsh & Funnell, 1995; Howard, 1995; Kremin et al., 2001). First, we used logistic regression to predict a correct versus an incorrect response. Second, we examined whether the production of semantic error responses was consistent across testing sessions by examining whether semantic errors produced in sessions 1 and 2 predicted the occurrence of the same error type in Session 3. Since the categorical variables "naming responses in Session 1/2/3" now included three rather than two categories (accurate response, semantic error, all other error types, including no specific responses), these analyses were run as multinomial logistic regressions (using the package "mlogit" in R (Croissant, 2013)).

C. In addition, McNemar's test for related samples was used to determine whether there was a difference in the consistency across sessions for inaccurate and accurate responses – were accurate responses and inaccurate responses equally consistent? McNemar's test was further used to compare the similarity in consistency between semantic errors and accurate responses and between semantic errors and other error types. The calculations were done using the package "exact2x2" in R (Fay, 2010) and Excel.

Homogeneity of the group

To examine whether the effects of consistency were homogeneous across the group we used linear mixed modelling using the *lme4* package in R (Bates, Maechler, Bolker, & Walker, 2015) and p-values were determined using the package *lmerTest* (Kuznetsova, Brockhoff, & Haubo, 2016). This allowed us to examine whether including naming responses from previous sessions as fixed effects in the group model significantly improved the model's fit – and thus, whether consistency was present at the group level. By including by-subject random slopes for all variables (naming responses from previous session, familiarity, age of acquisition, imageability, word frequency an length), we were able to assess the homogeneity of the group with respect to consistency measures and effects of each psycholinguistic variable: If including a random slope for a variable lead to significant improvement of the model compared to a reduced model without that random slope, participants were not homogenous with respect to the effects of that variable.

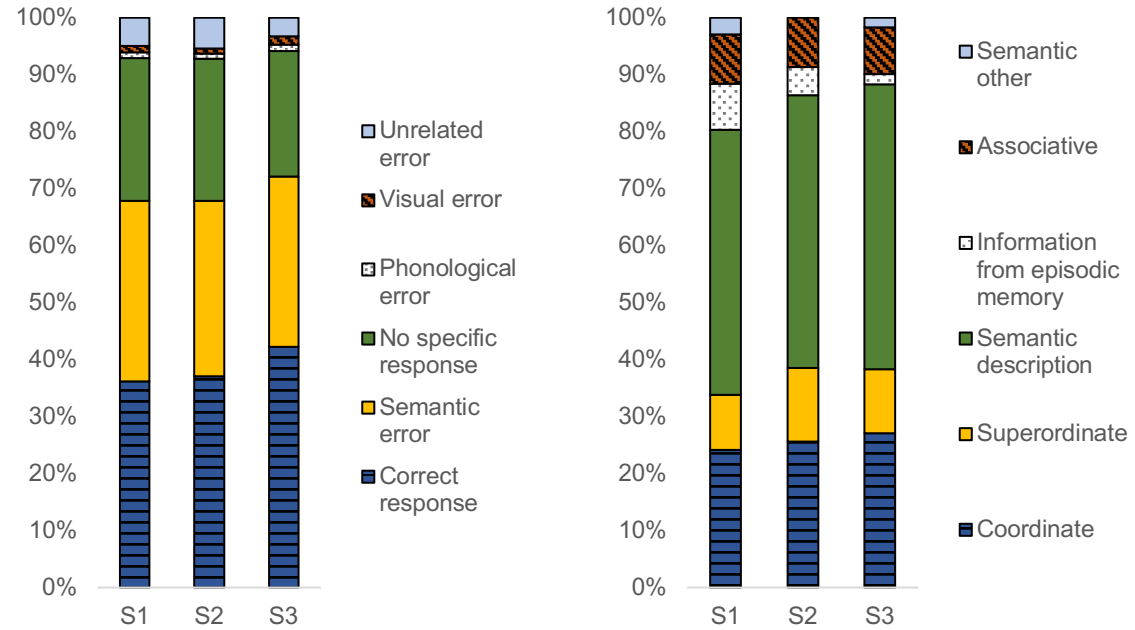
For all generalised linear models (i.e, the logistic and multinomial regressions and the mixed models), the continuous variables (familiarity, age of acquisition, imageability and frequency) were centred.

Results

Descriptive Error Analysis

Figure 2 shows the pattern of errors across the group as a whole, with individual patterns for each participant shown in Appendix C: Semantic errors were the most common

error type (31% of responses (SD= 14%), 50% of errors (SD =22%)) (see Figure 2a). When examining subtypes of semantic errors, semantic descriptions (48% of semantic errors, SD = 33%) and coordinate errors (26%, SD = 17%) were most frequent (Figure 2b). No (specific) responses (i.e., omissions) were the second most common error type across all sessions (mean 24% of responses (SD = 17%,); 39% of errors (SD= 27%). Accurate responses tended to increase from session to session for the group and for each individual. Accurate responses included those considered acceptable alternatives or part responses, as well (see above). Phonological errors (1% of responses (SD = 1%); 2% of errors (SD = 2%)) combined with visual errors (1% of responses (SD = 1%); 2 % of errors (SD = 1%)) and unrelated errors (5% of responses (SD = 5%); 7% of errors (SD = 9%)) form the smallest group of naming errors. This is consistent with the language assessment scores and the general disease pattern in svPPA, where phonological processing is usually spared.



a b

Figure 2. Average proportion of different response types in naming across sessions. Panel 2a: Major response types in each session. Panel 2b: Average proportion of different subtypes of semantic errors in each session. "Semantic other" included 1 subordinate error and 14 instances of unclassifiable semantically related responses.
S1= Session 1; S2 = Session 2; S3 = Session 3.

For those two participants where Sessions 2 and 3 were not consecutive baseline sessions due to technical problems in the recordings (participants G2 and K1), the increase in accuracy from Session 1 to Session 3 was significantly higher than in those participants where Sessions 1, 2 and 3 were consecutive baseline sessions (mean increase accuracy participants G2 and K1: 18, other participants: 3; Wilcoxon rank sum test W = 16, p = .049). This suggests a greater practice effect in those two participants as a result of the intervening session.

Statistical Analyses – single cases

Cohen's Kappa. A first analysis compared accuracy scores, that is, incorrect vs. correct answers, between the three sessions. As shown in Table 4 (left-hand column), all participants showed strong to fair consistency in their accuracy scores with four participants showing substantial strength of agreement in accuracy across sessions, five participants moderate strength of agreement, and one participant (G2[1]) fair strength of agreement. For J1, for example, there was a Kappa coefficient of 0.7 between the three sessions, reflecting very high consistency in accuracy across Session 1, 2 and 3. Participants J3, J2, G1 and J1 were the most consistent in their accuracy scores. When correlating naming severity (based on the SYDBAT naming scores) of participants with their Kappa values on accuracy consistency, there was no significant relationship (Pearson's r = 0.061, p = .867 (two-tailed)).

[1]G2 is an exception as she performed particularly poorly in Session 2 compared to Sessions 1 and 3, that is, very inconsistently. A closer look at her naming data reveals that she took a long time to respond to a picture in Session 2. This led to more responses being cut off by the 10 s time limit, which therefore had to be coded as no responses. G2's third naming session was not consecutive to Session 2, and perhaps the intervening sessions helped to increase her response speed so that more responses fell within the time limit.

Table 4: *Kappa values showing consistency over time in accuracy, in the four main response types (correct responses, semantic errors, no specific responses and other error responses) and eight response subtypes (correct response, other error response, coordinate, superordinate, associative, information from episodic memory, semantic descriptions or semantic other).*

	Consistency in accuracy		Consistency across 4 response types		Consistency across all 8 error subtypes	
Participants*	Kappa	Strength of agreement**	Kappa	Strength of agreement	Kappa	Strength of agreement
J3	0.625	substantial	0.405	fair	0.317	fair
T4	0.555	moderate	0.500	moderate	0.487	moderate
C2	0.600	moderate	0.558	moderate	0.467	moderate
G2	0.252	fair	0.197	slight	0.148	slight
K1	0.428	moderate	0.313	fair	0.330	fair
J2	0.687	substantial	0.561	moderate	0.539	moderate
G1	0.633	substantial	0.621	substantial	0.492	moderate
B1	0.521	moderate	0.337	fair	0.287	fair
S1	0.462	moderate	0.345	fair	0.338	fair
J1	0.700	substantial	0.636	substantial	0.584	moderate
mean	0.546		0.447		0.399	
SD	0.137		0.149		0.136	

* In this table and all following tables, participants are ordered by severity (from most severe to least severe as measured by accuracy of SydBat naming).
** The cut-off values for strength of agreement according to Landis and Koch (1977) are: < 0.00 (poor), 0.00-0.20 (slight), 0.21-0.40 (fair), 0.41-0.60 (moderate), 0.61-0.80 (substantial), 0.81-1.00 (almost perfect)

A second Kappa coefficient for each participant was calculated to assess consistency across correct responses, semantic errors, no specific responses and other error responses (including phonological, visual and unrelated errors; Table 4 middle column).

J1 and G1 were again most consistent. Overall, all ten participants were less consistent than they were when only accurate and inaccurate responses were compared. This difference was especially prominent for J3 and B1. In a final analysis using Kappa, we examined whether there was consistency in 8 different response types focusing on the different semantic error subtypes: accurate response, coordinate, superordinate, associative, information from episodic memory, semantic descriptions, semantic other and all other (nonsemantic) error responses. The results of the Kappa analyses show that consistency was still relatively high, as can be seen in Table 4 (righthand column). Those participants (T4, C2, J2, G1, J1) who had

shown the highest consistency in the previous analysis distinguishing between the four major response types (correct responses, semantic errors, no specific responses and other error types, see Table 4 middle column) showed moderate consistency and only a minimal reduction of the Kappa coefficients (by about .04 on average) in the more specific analysis including semantic error subtypes.

The consistency of semantic errors is further illustrated in a schematic overview of the semantic errors that were produced by each participant (see Figure 3). The two columns show the proportions of items that were semantic errors in Session 1 and also resulted in semantic errors in Session 2 or in Session 3 respectively. Figure 3 further shows if a response changed into an accurate response, a no specific response or another error response. For the majority of participants, most semantic error responses remained semantic errors in the following sessions rather than turning into any other response type (58% on average (SD=18%)). However, for participant S1, for example, almost 50% of the semantic errors in Session 1 became accurate responses in Session 2 and 3, and correspondingly, S1's Kappa value for the four different response types is only fair (0.345). For G1, on the other hand, 80% of semantic errors remained semantic errors in the following sessions and only very few changed to accurate responses or another error type. G1's Kappa value is consequently substantial (0.621).

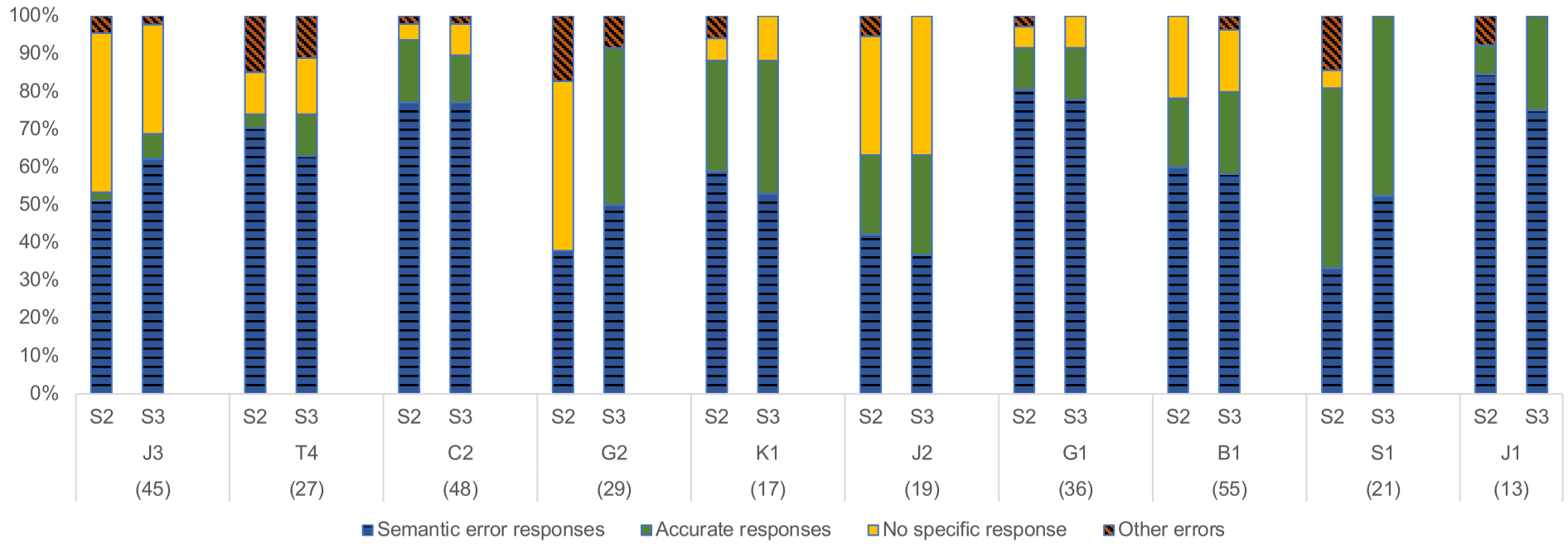


Figure 3. The proportions of semantic errors in Session 1 (number of items in brackets) that were produced as semantic errors, accurate responses, no specific responses or other error types in Sessions 2 and 3.

Relative consistency of response types – McNemar's test. The consistency analysis was supplemented by using McNemar's test to assess whether any response type was more or less consistent than another.

We first compared accurately and inaccurately named items, as shown in Table 5 (upper panel). If $p \geq .05$, then inaccurate and accurate responses were equally consistent or equally inconsistent across the sessions.

Comparing the consecutive sessions, Session 2 and 3, the majority of participants (8/10 participants) showed no difference in consistency between accurate and inaccurate items, except for G2 who was significantly more consistent in her accurate than inaccurate responses (more inaccurate responses became accurate responses than vice versa; $p<.001$; see Descriptive Error Analysis above) and J4 who showed a trend in the same direction ($p = .070$). When comparing the non-consecutive sessions (sessions 1 and 3), more participants showed differences in consistency between accurate and inaccurate items: three participants were significantly more consistent in their accurate compared to inaccurate responses (G2 $p=.001$; K1 $p=.001$; J2 $p=.049$) and the remaining three participants (T1: $p = .096$, B1: $p = .093$ and J1: $p = .070$) showed a trend in the same direction.

Table 5: *McNemar's tests examining relative consistency of correct and incorrect responses and semantic errors and no responses.*

Participants	**J3**		**T4**		**C2**		**G2**		**K1**		**J2**		**G1**		**B1**		**S1**		**J1**	
Sessions	**1 - 3**	**2 - 3**	**1 - 3**	**2 - 3**	**1 - 3**	**2 - 3**	**1 - 3**	**2 - 3**	**1 - 3**	**2 - 3**	**1 - 3**	**2 - 3**	**1 - 3**	**2 - 3**	**1 - 3**	**2 - 3**	**1 - 3**	**2 - 3**	**1 - 3**	**2 - 3**
correct and incorrect responses																				
No. of items in analysis	**100**		**124**		**106**		**86**		**94**		**103**		**102**		**95**		**90**		**79**	
consistent correct (in %)	9%	9%	12%	14%	33%	32%	21%	7%	26%	39%	35%	42%	38%	37%	16%	23%	41%	43%	72%	72%
consistent incorrect	80%	85%	73%	74%	49%	48%	55%	57%	37%	35%	49%	45%	41%	45%	60%	60%	31%	34%	18%	18%
inconsistent correct (changed from correct to incorrect)	7%	2%	4%	3%	9%	10%	3%	1%	7%	10%	4%	8%	13%	9%	7%	7%	13%	10%	1%	1%
inconsistent incorrect (changed from incorrect to correct)	4%	4%	10%	9%	8%	9%	21%	35%	30%	16%	13%	6%	8%	9%	17%	9%	14%	12%	9%	9%
exact p-values (two-tailed)	.549	.688	.096	.118	1.00	1.00	**.001**	**< .001**	**.001**	.307	**.049**	.791	.383	1.00	.093	.804	1.00	.824	.070	.070
Semantic errors and no specific responses (omissions).																				
No. of items in analysis	**75**	**80**	**64**	**70**	**49**	**48**	**42**	**41**	**32**	**27**	**46**	**44**	**39**	**42**	**50**	**50**	**50**	**50**	**11**	**10**
consistent semantic error (in %)	37%	28%	27%	36%	76%	85%	14%	10%	28%	59%	15%	14%	72%	71%	60%	48%	22%	26%	82%	90%
consistent no specific response	33%	43%	56%	53%	6%	6%	40%	56%	28%	30%	67%	66%	21%	21%	16%	14%	2%	4%	0%	0%
inconsistent (changed from semantic error to no specific response)	17%	9%	6%	7%	8%	2%	40%	20%	6%	7%	15%	18%	8%	5%	16%	18%	0%	0%	0%	0%
inconsistent (changed from no specific response to semantic error)	12%	21%	11%	4%	10%	6%	5%	15%	38%	4%	2%	2%	0%	2%	8%	20%	6%	8%	18%	10%
exact p-values (two-tailed)	.523	.064	.549	.727	1.00	.625	**< .001**	.791	**.013**	1.00	.070	**.039**	.250	1.00	.388	1.00	.250	.125	.500	1.00

In summary, the majority of the participants were equally consistent in their accurate and inaccurate responses across consecutive sessions. Where participants showed differences in consistency, inaccurate responses were always less consistent than accurate responses.

The same analysis was run examining only responses with semantic errors or omissions (no specific responses), to see which of these error types was more consistent or whether they were equally consistent for the individual participants. All items with accurate responses or other error types were excluded from this analysis. Results are summarised in Table 5 (lower panel).

In the consecutive sessions 2 and 3, all but one participant exhibited no significant difference, whereby semantic errors were as consistent (or inconsistent) as omissions, while for J3 semantic errors were less consistent than omissions. In the non-consecutive sessions 1 and 3, eight out of ten participants were equally consistent in their production of semantic errors compared to other error types, while two participants showed a significant difference (G2: $p < .001$ and K1: $p = .013$). For K1 semantic errors were more consistent, while for G2 semantic errors were less consistent compared to omissions.

Predicting naming responses from other variables – binary and multinomial logistic regressions. This analysis aimed to examine the factors predicting an accurate response in Session 3 (compared to any other response type) using logistic regression. We first added the psycholinguistic variables (familiarity, age of acquisition, imageability, frequency and length) as independent variables predicting Session 3 accuracy. Then, in two supplementary models we added either Session 2 accuracy or Session 1 accuracy as additional predictors, and finally we compared the predictive value of the different models. The full model outcomes are presented in Appendix D, and model comparisons are shown in Table 6.

For all participants, the accuracy of an item in a previous session was found to be a significant predictor of item accuracy in Session 3 and hence to improve the statistical model (Models 2a and 2b) over and above the predictive value of other language variables (Model 1;

see Table 6 below). The psycholinguistic variables (familiarity, age of acquisition, imageability, frequency and number of phonemes) differed in how far they predicted naming accuracy across participants. Familiarity, imageability and frequency were each significant predictors for three of the ten participants, age of acquisition (AoA) for two participants and length for one participant (Appendix D). Collinearity between predictor variables was accounted for by calculating the Variable Inflation Factor (VIF) for each predictor in each participant's regression model. Moreover, the correlation between predictor variables was determined individually for each participant since the picture set and therefore the values for language variables that were included in the models differed between participants. The results of these calculations are summarized in Appendix E. After careful inspection, we do not see cause for concerns about multicollinearity in our data, since both measures are below the values suggested as critical for multicollinearity by statisticians[2] (VIF<10; Field, Miles, & Field, 2012; Myers, 1990; r<0.8; Hutcheson & Sofroniou, 1999).

[2] For participants T4 and G1, familiarity and imageability correlated at rho = .850 and rho = .901, respectively. However, these participants did not perform any differently in the regression models. We therefore saw no concern in this relatively high correlation.

Table 6: *Model comparisons for predicting accuracy and semantic errors in Session 3.*

Participants	**J3**		**T4**		**C2**		**G2**		**K1**		**J2**		**G1**		**B1**		**S1**		**J1**	
No. of items	**100**		**124**		**106**		**86**		**94**		**103**		**102**		**95**		**90**		**79**	
	X^2	**p**	**X^2**	**p**	**X^2**	**p**	**X^2**	**p**	**X^2**	**p**	**X^2**	**p**	**X^2**	**p**	**X^2**	**p**	**X^2**	**p**	**X^2**	**p**
Model comparisons – *Logistic Regression models predicting accuracy in Session 3*																				
Model 1 – Model 2a[a]	16.878	**<.001**	22.080	**<.001**	36.128	**<.001**	6.881	**.009**	5.302	**.021**	40.361	**<.001**	17.562	**<.001**	4.188	**.041**	6.499	**.011**	38.582	**<.001**
Model 1 – Model 2b[b]	28.971	**<.001**	34.393	**<.001**	31.808	**<.001**	9.339	**.002**	19.230	**<.001**	40.441	**<.001**	21.489	**<.001**	15.430	**<.001**	6.499	**.011**	38.011	**<.001**
Model comparisons – *Multinomial Regression models predicting semantic errors in Session 3*																				
Model 1 - Model 2a	33.964	**<.001**	37.586	**<.001**	41.237	**<.001**	14.634	**<.001**	9.200	**.002**	51.313	**<.001**	41.581	**<.001**	16.787	**<.001**	24.709	**<.001**	42.838	**<.001**
Model 1 - Model 2b	41.210	**<.001**	71.035	**<.001**	61.925	**<.001**	11.924	**.001**	40.591	**<.001**	52.282	**<.001**	44.593	**<.001**	16.818	**<.001**	26.340	**<.001**	51.964	**<.001**

[a] **Model 2a** includes the independent variables familiarity, AoA, imageability, frequency, length and accuracy in Session 1.
[b] **Model 2b** includes the independent variables familiarity, AoA, imageability, frequency, length and accuracy in Session 2.

Multinomial regressions. Finally, as for the accuracy analyses earlier, a series of logistic regression analyses was carried out to reveal whether the production of semantic error responses was consistent across testing sessions by examining whether semantic errors produced in sessions 1 and 2 predicted the occurrence of the same error type in Session 3. The full model results are presented in Appendix F, and the model comparisons are shown in Table 6. We first examined whether a semantic error rather than an accurate response in Session 1 (2a) or 2 (2b) predicted a semantic error in Session 3 (Analysis 1, Appendix F, Upper panel), and, second, whether a semantic error vs any other error in Session 1 or Session 2 predicted a semantic error in Session 3 (Analysis 2, Appendix F, lower panel).

As in the logistic regression analyses, no language variable (familiarity, frequency etc.) consistently predicted naming responses of the third session significantly for all participants. However, for all participants a model including naming responses of Session 1 or 2 was always a significantly better fit than the model that contained only the language variables (Table 6).

1) Predicting a semantic error compared to an accurate response in Session 3

The likelihood of a semantic error in Session 3 rather than an accurate response was significantly increased by a semantic error compared to an accurate response in Session 1 and Session 2 for six participants (an Odds Ratios significantly >1 for participants J3, C2, K1, J2, G1, and T4 for whom in Session 2 the Odds ratio was uninterpretable) and for one further participant (B1) there was a significant effect in Session 2 and a trend in Session 1. A semantic error compared to another error type in Session 1 or 2 did not increase the likelihood of a semantic error rather than an accurate response in Session 3, except for participant G1 for Session 1.

2) Predicting a semantic error compared to other error types in Session 3

Whether a naming response in Session 3 was more likely to be a semantic error rather than an “other error” was significantly predicted by a semantic error versus another error type

in Session 1 or Session 2 for four participants (J3, T4, J2, G1), for one (B1) in Session 1 and not Session 2, and for two participants (C2, K1), in Session 2 but not Session 1 (C2 showed a trend in Session1). When they were contrasted with an accurate response, they did not increase the likelihood of a semantic error outcome in Session 3, except for one participant (C2, in Session 2).

Statistical Analyses - Group Analyses

Homogeneity of the group with respect to consistency was tested using GLM mixed modelling, removing one random effect variable at a time and comparing it to the full model (e.g., full model for naming accuracy in Session 3: glmer(Accuracy_S3~Accuracy_S1+Accuracy_S2+AoA+Imageability+Familiarity+Frequency+Length+(1+Accuracy_S1+Accuracy_S2+AoA+Imageability+Familiarity+Frequency+Length|participants), family="binomial", control=glmerControl(optimizer="bobyqa"), data = lme_model_data)). There was no significant difference between the models (all $p > .137$), confirming that the individuals with svPPA were homogenous with respect to consistency in their accuracy between sessions, and that there was no variability in the effects of the different psycholinguistic variables between the participants.

Consistency Analysis Summary

The results of this study showed that all of the individual participants (except for G2) were very consistent in their naming accuracy, with no evidence for any differences in the consistency shown by the participants[3]. Pictures that were named accurately (or inaccurately) in one session were likely to be named accurately (or inaccurately) in another session. This consistency in naming accuracy was independent of the severity of the language impairment. Including naming accuracy from Session 1 and 2 always resulted in a better prediction of naming responses in Session 3 than other psycholinguistic variables alone (e.g., familiarity,

[3] This includes participant K1, whose naming sessions were not consecutive – his pattern of consistency did not differ in any obvious way from that of the other participants.

frequency etc.). The majority of participants were as consistent in their accurate as in their inaccurate responses. In the few cases where a difference was measured, accurate responses were always more consistent than inaccurate responses which was most likely to be due to a practice effect.

The results for semantic errors and semantic error subtypes yielded consistency values ranging from slight to substantial with half the participants showing substantial to moderate consistency. For the majority of participants, multinomial regression analyses of error type consistency showed that a semantic error in Session 1 or Session 2 significantly predicted that the outcome of Session 3 would also be a semantic error instead of an accurate response or another error type. When comparing the consistency of the different error types, the majority of participants were equally consistent in the production of semantic errors and other error types.

Discussion

The current study investigated language production in the semantic variant Primary Progressive Aphasia (svPPA), which is assumed to be due to a central, amodal semantic impairment (Hodges et al., 1995). We investigated one characteristic of this semantic deficit, namely performance consistency on a picture naming task. Response consistency has been argued to be one of the main characteristics that can be used to distinguish between a "storage deficit" (where semantic representations are degraded and eventually lost) and an "access deficit" (where the representations remain unimpaired but cannot be accessed reliably) as the basic underlying deficits of the semantic system (Warrington & Cipolotti, 1996; Warrington & Shallice, 1979). People with svPPA are assumed to suffer from a storage deficit at the semantic level, with a progressive degradation of semantic representations (Hodges et al., 1995; McCarthy & Warrington, 2016; Mirman & Britt, 2014). Following this assumption, people with svPPA should perform highly consistently in tasks requiring semantic knowledge.

Our analyses revealed that, except for one individual, all the participants with svPPA were moderately to highly consistent in their naming accuracy. However, some participants were more likely to change their performance on items that were named inaccurately to accurate responses while showing more consistent performance on items that are already accurately named. Consequently, for these participants the number of accurate responses increased with each session, which can, most likely, be attributed to a practice or priming effect (for a detailed study on practice effects see, e.g., Nickels, 2002). Of course, testing could have also increased the participant's attention to these words resulting in attempts to use external sources to find and 'learn' the answers between sessions.

Indeed, for two participants, G2 and K1, where testing sessions were not entirely consecutive, but some sessions intervened between sessions 2 and 3, correct responses increased more in Session 3 than for the other participants. The greater spacing of their sessions meant that G2 and K1 may have had more opportunity for priming or to learn some of the previously incorrectly named items in the intermediate sessions.

High consistency was also found in the analysis of semantic errors, the most characteristic overt error type in svPPA. This showed that the majority of items that were named with a semantic error by the participants in the first session were again named with a semantic error in the following sessions.

Taken together, our results support the assumption that performance consistency is a characteristic of a semantic storage deficit, such as that exhibited by people with svPPA. In contrast to a storage impairment, an access impairment would predict variability in that sometimes a representation can be accessed and thus named correctly and sometimes not (McCarthy & Kartsounis, 2000). Some authors have cautioned that naming could be consistent even in an access impairment because of the impact of stimulus-related psycholinguistic variables such as word frequency, whereby, for example, low frequency words can be consistently inaccessible (e.g., Howard, 1995; Rapp and Caramazza (1993)). In

the present study we found that there was consistency over and above the effect of stimulus-related variables: Regressions showed that previous naming accuracy (or error type) reliably predicted subsequent responses even when these variables were included in the analysis.

By providing the first examination of consistency over multiple repetitions of the same stimuli in the same task, our results both support and extend the behavioural results of previous studies with people with svPPA (e.g., Budd et al., 2010; Hodges et al., 1995; Laisney et al., 2011). Moreover, they increase our understanding of how semantic features that represent semantic information (as proposed by theories of language production incorporating decomposed semantic representations (e.g., by Dell et al., 1997)) deteriorate in svPPA (Hodges et al., 1995). One might assume that a loss of semantic features is not compatible with naming consistency: this pattern seems to be more easily explained as a result of the loss of holistic concepts that are consequently consistently unavailable. Nonetheless, we would suggest that despite the high levels of consistency in our participants, the overall pattern is, in fact, more plausibly explained within a decomposed theory. First, there was not absolute consistency (as reflected by the Kappa values) and second, some participants showed practice effects across the three sessions. The way test items were selected for each participant may have contributed to this pattern: words were chosen if the person still had some semantic knowledge remaining. Given the partial remaining semantic knowledge, these items are more likely to show 'access-like' patterns (see Wilson et al., 2017, for evidence that some people with svPPA show access impairments). The loss of some semantic features leads to reduced (or inaccurate) activation of the lexical item. Activation of this lexical representation leads to gradual priming and greater accessibility on repeated presentation (Nickels, 2002). However, as semantic loss increases, activation reduces to the point that the lexical representation can never be retrieved. Hence, gradual loss of semantic features leads to initial inconsistency of access followed by later consistent inability to retrieve the form. In contrast, loss of a holistic semantic representation would be less likely to

result in a phase of inconsistent access to lexical forms - once semantics is lost, access to the lexical form is lost.

Our results showed that response variability was highest within the different subtypes of semantic error, suggesting that variability remained as to which features could still be activated to provide semantic information about the item. Once again, this points towards an intermediate stage of feature availability, which varies between participants but in all cases results in some variability of naming responses across sessions. For example, if some of the features that are necessary to name a cucumber are lost, a person with svPPA might be more likely to name it 'tomato' than 'lettuce' if he/she has recently (between sessions) more often thought of, talked about, or eaten a tomato. Future research could explore this by determining "individual exposure" or "individual familiarity" with the test items, or manipulating this experimentally.

Implications for future research and treatment

One question posed in the Introduction to this study was whether consistency could really be used as a criterion to distinguish a "storage" deficit from an "access" deficit. We have demonstrated that performance is indeed highly consistent in participants with svPPA, as predicted for an underlying storage deficit. We have argued that the consistency of naming performance found in the case series described here informs us about the consistent availability of necessary semantic features that lead to accurate naming, or their consistent unavailability leading to inaccurate naming. Conversely, the consistency or inconsistency of a participant's naming performance can allow us to draw conclusions about the extent to which particular lexical items are affected by the loss of semantic information. This could be an important insight for the choice of an appropriate treatment method. Recent literature reviews by Carthery-Goulart et al. (2013) and Jokel, Graham, Rochon, and Leonard (2014) on interventions in PPA have stressed that most interventions for individuals suffering from svPPA have focused on relearning of lost words. Importantly, participants relearn words more

effectively when they still have some retained semantic knowledge about the item (see, e.g., Jokel, Rochon, & Leonard, 2006). As we have suggested above, greater availability of semantic information in the form of semantic features is related to lower consistency in naming, especially within the semantic error subtypes. Consequently, for participants who show lower consistency, word-retrieval training might improve the ability to connect the remaining semantic information with its correct lexical label and thereby increase accuracy and reduce variability in naming. In contrast, for participants that show high consistency in naming errors and semantic error subtypes, treatment focusing on semantic feature generation might be more appropriate, thereby improving access to semantic features that have been lost due to deterioration (e.g. using approaches such as conceptual enrichment (COEN) therapy – Suárez-González et al, 2015; Suárez-González, Savage, & Caine, 2016).

Conclusion

The aim of this study was to clarify the significance of one characteristic of a storage deficit, which has been proposed to be the underlying deficit in svPPA: we focused on the investigation of performance consistency in picture naming to establish its role for the evaluation of the nature of semantic breakdown in this disease. Previous studies have only compared performance consistency across different tasks, between non-consecutive sessions, or looked only at the consistency in accuracy (e.g., Coccia et al., 2004; Hodges et al., 1995; Jefferies & Lambon Ralph, 2006). In this study we found further evidence of consistency in performance at an item level in naming accuracy and also for the production of semantic errors independent of word frequency, familiarity or other item-related variables. These results provide further insights into the structure of semantic representations and the way they can be damaged in a neurodegenerative disease like svPPA. Specifically, our results suggest that consistency varies as a function of semantic feature availability in svPPA and that this availability is not strictly related to disease severity. We suggest that information about consistency in naming performance cannot only be taken to diagnose and thus localise the

underlying impairment (semantic storage vs. access deficit) but also to inform and thus select an appropriate language treatment technique.

References

Acosta-Cabronero, J., Patterson, K., Fryer, T. D., Hodges, J. R., Pengas, G., Williams, G. B., & Nestor, P. J. (2011). Atrophy, hypometabolism and white matter abnormalities in semantic dementia tell a coherent story. *Brain, 134*(7), 2025-2035. doi:10.1093/brain/awr119

Alario, F.-X., & Ferrand, L. (1999). A set of 400 picutres standardized for French: Norms for name agreement, image agreement, familiarity, visual complexity, image variablitly, and age of acquisition. *Behavior Research Methods, Instruments, & Computers, 32*, 531-552.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi:10.18637/jss.v067.i01

Best, W., Nickels, L., & Williamson, L. (2005). *A coding system for picture naming responses with a focus on semantic errors*. Unpublished manuscript.

Bishop, D. V. M. (1989). *The Test for Reception of Grammar*. London: Medical Research Council.

Bozeat, S., Lambon Ralph, M. A., Patterson, K., Garrard, P., & Hodges, J. R. (2000). Non-verbal semantic impairment in semantic dementia. *Neuropsychologia, 38*(9), 1207-1215. doi:10.1016/S0028-3932(00)00034-8

Brambati, S. M., Rankin, K. P., Narvid, J., Seeley, W. W., Dean, D., Rosen, H. J., . . . Gorno-Tempini, M. L. (2009). Atrophy progression in semantic dementia with asymmetric temporal involvement: A tensor-based morphometry study. *Neurobiology of Aging, 30*(1), 103-111. doi:10.1016/j.neurobiolaging.2007.05.014

Budd, M. A., Kortte, K., Cloutman, L., Newhart, M., Gottesman, R. F., Davis, C., . . . Hillis, A. E. (2010). The nature of naming errors in primary progressive aphasia versus acute post-stroke aphasia. *Neuropsychology, 24*(5), 581-589. doi:10.1037/a0020287

Buzsáki, G. (2010). Neural Syntax: Cell Assemblies, Synapsembles, and Readers. *Neuron, 68*(3), 362-385. doi:10.1016/j.neuron.2010.09.023

Carthery-Goulart, M. T., Silveira, A. d. C. d., Machado, T. H., Mansur, L. L., Parente, M. A. d. M. P., Senaha, M. L. H., . . . Nitrini, R. (2013). Nonpharmacological interventions for cognitive impairments following primary progressive aphasia: A systematic review of the literature. *Dementia & Neuropsychologia, 7*, 122-131.

Chertkow, H., & Bub, D. (1990). Semantic memory loss in dementia of Alzheimer's type. *Brain, 113*, 397-417.

Coccia, M., Bartolini, M., Luzzi, S., Provinciali, L., & Lambon Ralph, M. A. (2004). Semantic memory is an amodal, dynamic system: Evidence from the interaction of naming and object use in semantic dementia. *Cognitive Neuropsychology, 21*(5), 513-527. doi:10.1080/02643290342000113

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Coughlan, A. K., & Warrington, E. K. (1981). The impairment of verbal semantic memory: a single case study. *Journal of Neurology, Neurosurgery & Psychiatry, 44*(12), 1079.

Croissant, Y. (2013). Estimation of multinomial logit models in R : The mlogit Packages. Retrieved from https://cran.r-project.org/web/packages/mlogit/index.html

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexial access in aphasic and nonaphasic speakers. *Psychological Review, 104*, 801-838.

Fay, M. P. (2010). Two-sided Exact Tests and Matchin Confidence Intervals for Discrete Data. *R Journal, 2*(1), 53-58.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Los Angeles: SAGE Publications.

Gilhooly, K. J., & Hay, D. (1977). Imagery, concreteness, age-of-acquisition, familiarity, and meaningfulness values for 205 five-letter words having single-solution anagrams. *Behavior Research Methods & Instrumentation, 9*(1), 12-17. doi:10.3758/BF03202210

Gorno-Tempini, M. L., Hillis, A. E., Weintraub, S., Kertesz, A., Mendez, M., Cappa, S. F., . . . Grossman, M. (2011). Classification of primary progressive aphasia and its variants. *Neurology, 76*(11), 1006-1014. doi:10.1212/WNL.0b013e31821103e6

Gotts, S. J., & Plaut, D. C. (2002). The impact of synaptic depression following brain damage: a connectionist account of "access/refractory" and "degraded-store" semantic impairments. *Cognitive, Affective, & Behavioral Neuroscience, 2*(3), 187-213.

Hebb, D. O. (1949). *The organization of behaviour*. New York: John Wiley & Sons.

Hirsh, K. W., & Funnell, E. (1995). Those old, familiar things: Age of acquisition, familiarity and lexical access in progressive aphasia. *Journal of Neurolinguistics, 9*, 23-32.

Hodges, J. R., Graham, N., & Patterson, K. (1995). Charting the progression in semantic dementia: Implications for the organisation of semantic memory. *Memory, 3*, 463-495. doi:10.1080/09658219508253161

Hodges, J. R., Martinos, M., Woollams, A. M., Patterson, K., & Adlam, A.-L. R. (2008). Repeat and Point: Differentiating semantic dementia from progressive non-fluent aphasia. *Cortex, 44*(9), 1265-1270. doi:10.1016/j.cortex.2007.08.018

Hodges, J. R., Patterson, K., Oxbury, S., & Funnell, E. (1992). Semantic Dementia. Progressive fluent aphasia with temporal lobe atrophy. *Brain, 115*(6), 1783-1806. doi:10.1093/brain/115.6.1783

Howard, D. (1995). Lexical Anomia: Or the Case of the Missing Lexical Entries. *The Quarterly Journal of Experimental Psychology Section A, 48*, 999-1023.

Hutcheson, G. D., & Sofroniou, N. (1999). *The Multivariate Social Scientist: Introductory Statistics Using Generalized Linear Models*. London: SAGE Publications.

Jefferies, E., & Lambon Ralph, M. A. (2006). Semantic impairment in stroke aphasia versus semantic dementia: a case-series comparison. *Brain, 129*(8), 2132-2147.

Jokel, R., Graham, N. L., Rochon, E., & Leonard, C. (2014). Word retrieval therapies in primary progressive aphasia. *Aphasiology, 28*(8-9), 1038-1068. doi:10.1080/02687038.2014.899306

Jokel, R., Rochon, E., & Leonard, C. (2006). Treating anomia in semantic dementia: Improvement, maintenance, or both? *Neuropsychological Rehabilitation, 16*(3), 241-256. doi:10.1080/09602010500176757

Kremin, H., Perrier, D., Wilde, M. D., Dordain, M., Bayon, A. L., Gatignol, P., . . . Arabia, C. (2001). Factors predicting success in picture naming in Alzheimer's disease and primary progressive aphasia. *Brain and Cognition, 46*, 180-183. doi:10.1006/brcg.2000.1270

Kuznetsova, A., Brockhoff, P. B., & Haubo, R. (2016). lmerTest: Tests in Linear Mixed Effect Models. Retrieved from https://CRAN.R-project.org/package=lmerTest

Laisney, M., Giffard, B., Belliard, S., de la Sayette, V., Desgranges, B., & Eustache, F. (2011). When the zebra loses its stripes: Semantic priming in early Alzheimer's disease and semantic dementia. *Cortex, 47*(1), 35-46. doi:10.1016/j.cortex.2009.11.001

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics, 33*, 159-174.

Marczinski, C. A., & Kertesz, A. (2006). Category and letter fluency in semantic dementia, primary progressive aphasia, and Alzheimer's disease. *Brain and language, 97*(3), 258-265. doi:10.1016/j.bandl.2005.11.001

Marques, J. F., & Charnallet, A. (2013). The role of feature sharedness in the organization of semantic knowledge: insights from semantic dementia. *Neuropsychology, 27*(2), 266-274. doi:10.1037/a0032058

McCarthy, R. A., & Kartsounis, L. D. (2000). Wobbly words: Refractory anomia with preserved semantics. *Neurocase, 6*, 487-497. doi:10.1080/13554790008402719

McCarthy, R. A., & Warrington, E. K. (2016). Past, present, and prospects: Reflections 40 years on from the selective impairment of semantic memory (Warrington, 1975). *The Quarterly Journal of Experimental Psychology, 69*(10), 1941-1968. doi:10.1080/17470218.2014.980280

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience, 4*(4), 310-322.

Mesulam, M., Rogalski, E., Wieneke, C., Cobia, D., Rademaker, A., Thompson, C., & Weintraub, S. (2009). Neurology of anomia in the semantic variant of primary progressive aphasia. *Brain, 132*(9), 2553-2565. doi:10.1093/brain/awp138

Meyers, J. E., & Meyers, K. R. (1995). Rey complex figure test under four different administration procedures. *The Clinical Neuropsychologist, 9*(1), 63-67. doi:10.1080/13854049508402059

Mioshi, E., Dawson, K., Mitchell, J., Arnold, R., & Hodges, J. R. (2006). The Addenbrooke's Cognitive Examination Revised (ACE-R): a brief cognitive test battery for dementia screening. *International journal of geriatric psychiatry, 21*(11), 1078-1085. doi:10.1002/gps.1610

Mioshi, E., Hsieh, S., Savage, S., Hornberger, M., & Hodges, J. R. (2010). Clinical staging and disease progression in frontotemporal dementia. *Neurology, 74*(20), 1591-1597. doi:10.1212/WNL.0b013e3181e04070

Mirman, D., & Britt, A. E. (2014). What we talk about when we talk about access deficits. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 369*(1634).

Montembeault, M., Brambati, S. M., Joubert, S., Boukadi, M., Chapleau, M., Laforce, R., Jr., . . . Rouleau, I. (2017). Naming unique entities in the semantic variant of primary progressive aphasia and Alzheimer's disease: Towards a better understanding of the semantic impairment. *Neuropsychologia, 95*, 11-20. doi:https://doi.org/10.1016/j.neuropsychologia.2016.12.009

Mummery, C. J., Patterson, K., Price, C. J., Ashburner, J., Frackowiak, R. S. J., & Hodges, J. R. (2000). A voxel-based morphometry study of semantic dementia: Relationship between temporal lobe atrophy and semantic memory. *Annals of Neurology, 47*(1), 36-45. doi:10.1002/1531-8249(200001)47:1<36::AID-ANA8>3.0.CO;2-L

Myers, R. H. (1990). *Classical and modern regression with applications* (2nd ed.). Boston: Duxbury.

Nickels, L. (2002). Improving word finding: Practice makes (closer to) perfect? *Aphasiology, 16*, 1047-1060. doi:10.1080/02687040143000618

Patterson, K., & Hodges, J. R. (2000). Sementic Dementia: one window on the structure and organisation of semantic memory. In F. Boller & J. Grafman (Eds.), *Handbook of Neuropsychology* (2nd ed., Vol. 2, pp. 313-333). Amsterdam, Netherlands: Elsevier.

Pavio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness: Values for 925 nouns. *Journal of Experimental Psychology, Monograph Supplement, 1*(2), 1-25.

R Core Team. (2016). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rapp, B., & Caramazza, A. (1993). On the distinction between deficits of access and deficits of storage: A question of theory. *Cognitive Neuropsychology, 10*, 113-141.

Savage, S. A., Ballard, K. J., Piguet, O., & Hodges, J. R. (2013). Bringing words back to mind - Improving word production in semantic dementia. *Cortex, 49*(7), 1823-1832. doi:10.1016/j.cortex.2012.09.014

Savage, S. A., Hsieh, S., Leslie, F., Foxe, D., Piguet, O., & Hodges, J. R. (2013). Distinguishing subtypes in primary progressive aphasia: application of the Sydney language battery. *Dementia and geriatric cognitive disorders, 35*(3-4), 208-218. doi:10.1159/000346389

Savage, S. A., Piguet, O., & Hodges, J. R. (2014). Giving words new life: generalization of word retraining outcomes in semantic dementia. *Journal of Alzheimers Disease, 40*(2), 309-317. doi:10.3233/JAD-131826

Savage, S. A., Piguet, O., & Hodges, J. R. (2015). Cognitive intervention in semantic dementia: maintaining words over time. *Alzheimer Disease & Associated Disorders, 29*(1), 55-62. doi:10.1097/wad.0000000000000053

Shallice, T. (1987). Impairments of semantic processing: multiple dissociations. In M. Coltheart, G. Sartori, & R. Job (Eds.), *The cognitive neuropsychology of language* (pp. 111-127). London: L. Erlbaum Associates.

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary*: Oxford University Press.

Suárez-González, A., Heredia, C. G., Savage, S. A., Gil-Néciga, E., García-Casares, N., Franco-Macías, E., . . . Caine, D. (2015). Restoration of conceptual knowledge in a case of semantic dementia. *Neurocase, 21*(3), 309-321. doi:10.1080/13554794.2014.892624

Suárez-González, A., Savage, S. A., & Caine, D. (2016). Successful short-term re-learning and generalisation of concepts in semantic dementia. *Neuropsychological Rehabilitation*, 1-15. doi:10.1080/09602011.2016.1234399

Tombaugh, T. (2004). Trail Making Test A and B: Normative data stratified by age and education. *Archives of Clinical Neuropsychology, 19*(2), 203-214. doi:10.1016/s0887-6177(03)00039-8

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology, 67*(6), 1176-1190. doi:10.1080/17470218.2013.850521

Vigliocco, G., Vinson, D. P., Lewis, W., & Garrett, M. F. (2004). Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology, 48*(4), 422-488. doi:10.1016/j.cogpsych.2003.09.001

Warrington, E. K., & Cipolotti, L. (1996). Word comprehension: The distinction between refractory and storage impairments. *Brain, 119*, 611-625.

Warrington, E. K., & Shallice, T. (1979). Semantic access dyslexia. *Brain, 102*, 43-63.

Wechsler, D. (2008). *Wechsler adult intelligence scale–Fourth Edition (WAIS–IV)*. San Antonio, TX: NCS Pearson.

Wilson, S. M., Dehollain, C., Ferrieux, S., Christensen, L. E. H., & Teichmann, M. (2017). Lexical access in semantic variant PPA: Evidence for a post-semantic contribution to naming deficits. *Neuropsychologia, 106*(Supplement C), 90-99. doi:10.1016/j.neuropsychologia.2017.08.032

Appendix A

Stimuli included in the analysis (n=267)

airconditioner	clothes line	hanky	moth	remote control	swan
alpacas	coasters	hat	mouse	rhinoceros	t-shirt
aluminium foil	cockroach	heater	mushrooms	rubber band	tape measure
apple	coffee	hippopotamus	nail	safety pins	tea
apron	coffee grinder	hose	nectarines	salt and pepper	tea towel
asparagus	coffee machine	hyena	oats	saucepan	teabag
avocado	colander	ice cream	okapi	sausages	teapot
bacon	comb	iron	olive	scales	television
banana	computer	jacket	olive oil	scarf	termite
barbecue	corn	jam	onion	scissors	thimbles
bath	cow	jeans	orange	screw	thongs
beans	crocodile	jug	oven	screwdriver	tiger
beater	cucumber	juice	paintbrush	secateurs	tissues
bees	cup	jumper	paper clip	serviettes	toaster
beetle	dental floss	kangaroo	paper towel	sewing machine	toilet
belt	dish cloth	kettle	parsley	shampoo	toilet paper
blower	dishwasher	kiwi fruit	passionfruit	shaving cream	tomato sauce
bobbins	dragonfly	knife	pasta	shirt	tomatoes
bowl	dryer	knitting needles	peanuts	shorts	tongs
bread	duck	koala	pear	shower	toothbrush
broccoli	dustpan and brush	kookaburra	peas	silver beet	toothpaste
broom	edger	ladle	peeler	silverfish	towel
butter	eggplant	ladybird	pelican	singlet	trousers
butterfly	eggs	lathe	pen	sink	tweezers
button	elephant	lawn mower	pencil	skink	underpants
calculator	emu	leaf blower	penguin	slipper	vacuum cleaner
can opener	fabric	leeks	phone	snake	wallet
capsicum	fan	lemon	pig	sneakers	warthog
cardigan	fence	leopard	pineapple	snow peas	washing machine
carrot	fly	lettuce	placemats	soap	wasp
cashews	fork	lime	plate	socks	watch
casserole dish	fridge	lion	pliers	spade	watermelon
cauliflower	frog	lizard	plug	spaghetti	whales
celery	frying pan	loppers	polo shirt	spatula	whisk
centipede	galah	magpie	pot	spider	wildebeest
chainsaw	garlic	mandarin	potato masher	spoon	wine cask
cheese	garlic crusher	matches	potatoes	stapler	wine glass
Cheetah	giraffe	mayonnaise	power point	steak	wok
chicken	glad wrap	meerkat	praying mantis	stepladder	wombat
chilli	grapefruit	microwave	rabbit	sticky tape	yoghurt
chisel	grapes	milk	rainbow lorikeet	stove	zipper
chocolate	grasshopper	mince	raisins	strawberries	zucchini
chopping board	grater	mint	rake	sugar	
cicada	hammer	mirror	ravioli	sultanas	
cling wrap	hanger	mosquito	razor	sunglasses	

Appendix B

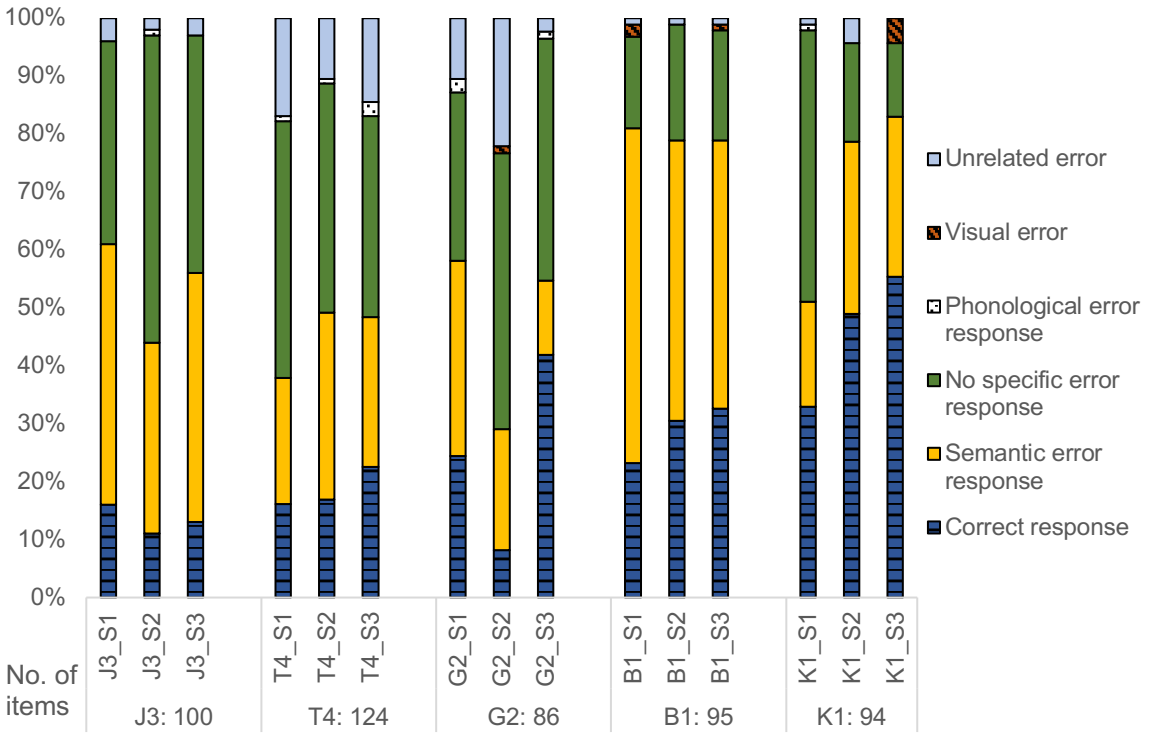
Stimuli excluded from the analysis or different target label chosen (n=51)

Very specific low-frequency items were excluded from the analyses completely. Where an item label was very specific (e.g., “electric frying pan”) or included two options (e.g., “beater or mixer”), it was replaced by the more common option. This then served as the target label and as the point of reference for error coding.

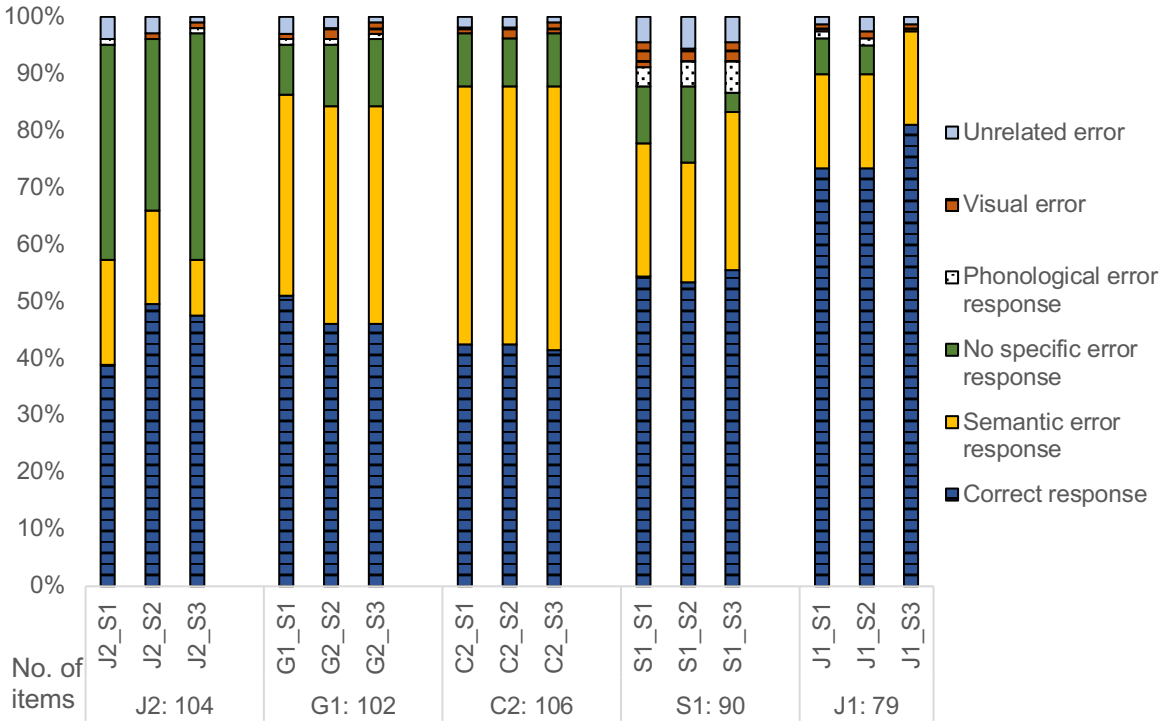
Excluded	**replaced by**	**Excluded**	**replaced by**
African buffalo	/	jandles	/
backing fabric	/	kitchen scales	scales
batting or wadding	/	leaf rake	rake
beater or mixer	beater	linen cupboard	cupboard
border	/	measuring tape	tape measure
brazil nuts	/	nailbrush	/
bull ant	/	office or study	office
ceiling fan	fan	peppers	capsicum
choc chip biscuits	/	pin cushions	/
circular knitting needles	knitting needles	pot or saucepan	pot
coat hanger	hanger	potato peeler	peeler
coffee cup or mug	mug	quilter's ruler	/
coffee percolator	/	razor or shaver	razor
collared shirt	/	rotary cutter	/
cooktop or stove	stove	rotary mat	/
crotchet hooks	/	sandals / thongs / flip flops	thongs
edge trimmer	edger	sewing needles	/
electric drill	/	sewing pins	/
electric frying pan	frying pan	sewing thread	/
face cloth	/	shaver	razor
fettuccine	/	shifting spanners	/
garden hose	hose	teatowel	tea towel
griller	/	snowpeas	snow peas
hacksaw	/	whipper snipper	/
hedger	/	wire cutters	/
icecream drumstick	/		

Appendix C

Descriptive Error Analyses for individual participants



Appendix Figure C1. Response patterns of participants with the fewest correct responses.



Appendix Figure C2. Response patterns of participants with the most correct responses.

Appendix D

Full model results of Logistic Regressions predicting accuracy in Session 3

Logistic Regression results showing the significance of the models predicting accuracy in Session 3.

Participants	**J3**		**T4**		**C2**		**G2**		**K1**		**J2**		**G1**		**B1**		**S1**		**J1**	
No. of items	**100**		**124**		**106**		**86**		**94**		**103**		**102**		**95**		**90**		**79**	
	Odds Ratio[a]	**p**	**Odds Ratio**	**p**	**Odds Ratio**	**p**	**Odds Ratio**	**p**	**Odds Ratio**	**p**	**Odds Ratio**	**p**	**Odds Ratio**	**p**	**Odds Ratio**	**p**	**Odds Ratio**	**p**	**Odds Ratio**	**p**
Model 1 – only language variables																				
Familiarity	4.163	**.038**	.943	.916	.391	**.013**	.804	.648	.656	.277	.620	.155	.912	.891	2.315	.084	.229	**.002**	.865	.727
AoA	2.676	.138	.925	.862	.665	.301	.193	**.010**	1.011	.978	.542	.126	.441	.106	.336	**.049**	.990	.978	1.416	.496
Imageability	1.201	.745	1.448	.437	2.104	.081	1.248	.673	2.495	**.015**	1.535	.255	3.699	**.037**	.613	.315	3.097	**.012**	2.248	.153
Frequency	2.805	.107	3.744	**.007**	1.045	.886	5.310	**.008**	1.042	.900	1.896	.128	1.582	.263	1.854	.194	2.800	**.008**	1.407	.358
Length	1.241	.275	1.317	.059	1.006	.963	1.389	.071	.933	.600	1.070	.589	1.372	**.040**	.908	.546	.966	.813	.973	.864
Model 2a[b] – including responses from Session 1																				
Accuracy S1	20.060	**<.001**	17.327	**<.001**	20.392	**<.001**	8.111	**.016**	3.460	**.026**	82.674	**<.001**	9.846	**<.001**	3.799	**.044**	3.951	**.012**	200.16	**<.001**
Model 2b[c] – including responses from Session 2																				
Accuracy S2	161.03	**<.001**	40.287	**<.001**	14.645	**<.001**	38.677	**.013**	8.989	**<.001**	32.307	**<.001**	12.756	**<.001**	10.995	**<.001**	12.533	**<.001**	184.03	**<.001**

[a] The likelihood of an outcome to be in the comparison rather than the reference category is expressed by Odds Ratios >1. Odds Ratios are the exponentiation of the regression coefficients for each variable.
[b] **Model 2a** includes the independent variables familiarity, AoA, imageability, frequency, length and accuracy in Session 1.
[c] **Model 2b** includes the independent variables familiarity, AoA, imageability, frequency, length and accuracy in Session 2.

Appendix E

Collinearity of psycholinguistic variables

Variable Inflation Index (VIF) for each participant for each logistic regression model

Participant		**J3**	**T4**	**C2**	**G2**	**K1**	**J2**	**G1**	**B1**	**S1**	**J1**
Variable	**Model**										
Familiarity	1	2.783	2.615	2.760	1.869	3.063	1.961	2.192	1.580	3.453	2.124
AoA	1	3.501	2.460	3.193	1.984	3.084	2.580	2.441	1.921	2.003	2.803
Imageability	1	1.842	2.045	2.856	1.806	2.488	2.154	1.758	1.974	2.860	3.769
Frequency	1	2.091	1.973	2.038	1.705	2.274	2.539	2.340	1.812	1.988	1.718
Length	1	1.545	1.728	1.489	1.839	1.730	1.644	1.942	1.415	1.425	1.460
Familiarity	2a	4.594	2.985	3.044	1.887	3.344	3.215	2.201	1.542	3.663	2.465
AoA	2a	6.273	2.868	3.459	2.144	3.181	2.115	2.704	1.988	2.085	3.978
Imageability	2a	2.471	2.387	3.134	1.820	2.578	2.319	1.820	1.959	3.148	5.246
Frequency	2a	2.542	2.182	1.994	1.899	2.302	2.462	2.338	1.949	2.140	1.765
Length	2a	1.655	1.939	1.488	2.121	1.779	1.724	1.982	1.533	1.488	1.543
Accuracy S1	2a	1.090	1.119	1.266	1.128	1.167	2.151	1.063	1.140	1.122	1.322
Familiarity	2b	3.176	2.779	3.046	1.717	3.413	2.124	2.282	1.602	4.013	1.547
AoA	2b	4.542	2.795	3.267	2.690	3.225	2.281	2.651	2.043	2.100	3.796
Imageability	2b	1.919	2.477	3.130	1.888	2.497	2.373	1.851	2.007	3.291	3.655
Frequency	2b	2.370	2.326	2.163	1.759	2.236	2.663	2.401	1.875	2.120	1.840
Length	2b	1.657	2.225	1.521	2.052	1.770	1.672	1.908	1.531	1.491	1.665
Accuracy S2	2b	1.622	1.168	1.137	1.394	1.182	1.175	1.090	1.108	1.137	1.318

Correlations between psycholinguistic variables for each participant

Participants	**J3**	**T4**	**C2**	**G2**	**K1**	**J2**	**G1**	**B1**	**S1**	**J1**
Correlations	**r**	**r**	**r**	**r**	**r**	**r**	**r**	**r**	**r**	**r**
Fam_AoA	-.745	-.734	-.709	-.665	-.706	-.573	-.774	-.598	-.640	-.493
Fam_Freq	.436	.500	.375	.335	.416	.457	.517	.375	.444	.147
Fam_Imag	.753	.850	.786	.755	.747	.695	.901	.781	.772	.630
Freq_AoA	-.600	-.651	-.630	-.602	-.647	-.693	-.668	-.600	-.582	-.491
Imag_AoA	-.664	-.691	-.743	-.692	-.642	-.694	-.746	-.698	-.596	-.779
Imag_Freq	.412	.509	.449	.405	.441	.529	.496	.473	.472	.417
Leng_AoA	.448	.342	.383	.487	.487	.450	.348	.275	.405	.107
Leng_Fam	-.273	-.039	-.192	-.279	-.206	-.145	-.056	-.023	-.153	.051
Leng_Freq	-.644	-.578	-.553	-.635	-.629	-.619	-.578	-.514	-.559	-.519
Leng_Imag	-.304	-.091	-.203	-.368	-.251	-.247	-.105	-.129	-.184	-.101

Fam = Familiarity, AoA = Age of Acquisition, Freq = Frequency, Imag = Imageability, Phon = Number of phonemes

Appendix F

Full model results of Multinomial Regressions predicting semantic errors in Session 3

Multinomial Regression analyses predicting accurate, semantic error or other errors in Session 3

Analysis 1: Predicting semantic errors vs accurate responses in Session 3

Participants	**J3**		**T4**		**C2**		**G2**		**K1**		**J2**		**G1**		**B1**		**S1**		**J1**	
No. of items	**100**		**124**		**106**		**86**		**94**		**103**		**102**		**95**		**90**		**79**	
	Odds Ratio[a]	**p**	**Odds Ratio**	**p**	**Odds Ratio**	**p**	**Odds Ratio**	**p**	**Odds Ratio**	**p**	**Odds Ratio**	**p**	**Odds Ratio**	**p**	**Odds Ratio**	**p**	**Odds Ratio**	**p**	**Odds Ratio**	**p**
Model 1 – only language variables																				
Familiarity	0.196	**.024**	1.153	.829	2.997	**.010**	2.995	.137	1.839	.179	0.947	.909	1.734	.462	0.402	.068	4.594	**.004**	1.014	.976
AoA	0.299	.084	0.811	.700	1.634	.243	11.254	**.003**	1.439	.421	1.260	.728	2.334	.118	2.914	.057	1.112	.789	0.494	.247
Imageability	0.809	.720	0.513	.264	0.428	.067	0.597	.481	0.426	**.045**	0.315	.068	0.191	**.017**	1.813	.238	0.292	**.014**	0.433	.188
Frequency	0.403	.170	0.128	**<.001**	0.874	.680	0.213	**.033**	1.194	.646	1.754	.441	0.505	.125	0.627	.344	0.373	**.020**	0.508	.112
Phonemes	0.852	.445	0.649	**.026**	0.973	.830	0.555	**.037**	1.038	.805	1.206	.374	0.717	**.050**	1.132	.447	0.978	.893	0.898	.559
Model 2a – including responses from Session 1 as predictors																				
Contrasts																				
Sem vs Acc in Session 1 (S1)	19.013	**.002**	26.144	**.005**	33.510	**<.001**	4.321	.232	5.841	**.023**	48.178	**.005**	14.873	**<.001**	3.505	.068	3.802	.058	NA	NA
Sem vs Oth S1	1.366	.811	3.646	.140	2.902	.209	7.438	.052	2.589	.150	6.441	.150	28.646	**.025**	2.573	.261	0.375	.248	9.518	.135
Model 2b – including responses from Session 2 as predictors																				
Contrasts																				
Sem vs Acc in Session 2 (S2)	70.198	**.002**	NA	NA	47.102	**<.001**	NA	NA	8.337	**.001**	39.380	**.006**	11.954	**<.001**	10.419	**.002**	42.165	**<.001**	NA	NA
Sem vs Oth S2	0.593	.662	2.386	2.386	12.670	.008	4.719	.141	5.563	.073	15.914	.069	2.739	.381	1.307	.771	4.115	.136	5.159	.182

[a] The likelihood of an outcome to be in the comparison rather than the reference category is expressed by Odds Ratios >1. Odds Ratios are the exponentiation of the regression coefficients for each variable. In some cases, Odds Ratio values could not be computed or were extremely high (<1000) when not enough values entered the models. These Odd Ratio values are marked "NA". They cannot be interpreted and were therefore ignored.

Continued from above.

Analysis 2: Predicting semantic errors vs other error types in Session 3

Participants	**J3**		**T4**		**C2**		**G2**		**K1**		**J2**		**G1**		**B1**		**S1**		**J1**	
No. of items	**100**		**124**		**106**		**86**		**94**		**103**		**102**		**95**		**90**		**79**	
	Odds Ratio	**p**	**Odds Ratio**	**p**	**Odds Ratio**	**p**	**Odds Ratio**	**p**	**Odds Ratio**	**p**	**Odds Ratio**	**p**	**Odds Ratio**	**p**	**Odds Ratio**	**p**	**Odds Ratio**	**p**	**Odds Ratio**	**p**
Model 1 – only language variables																				
Familiarity	0.674	.310	1.125	.807	1.890	.296	2.924	.099	1.604	.416	0.494	.174	4.121	.092	0.764	.571	1.121	.854	0.538	.538
AoA	0.655	.261	0.684	.347	1.372	.597	2.648	.112	2.721	.095	0.630	.479	1.060	.919	0.914	.824	1.301	.597	0.266	.235
Imageability	0.951	.882	0.671	.395	0.647	.517	0.706	.579	1.184	.763	0.410	.144	0.361	.161	1.566	.409	0.774	.660	0.971	.981
Frequency	1.250	.489	0.372	**.009**	0.592	.304	1.153	.755	1.712	.243	4.322	**.035**	0.387	.107	1.723	.156	1.152	.749	0.221	.108
Phonemes	1.112	.453	0.815	.199	0.904	.576	0.727	.194	0.920	.668	1.370	.142	0.919	.706	1.120	.475	0.870	.474	0.548	.125
Model 2a – including responses from Session 1 as predictors																				
Contrasts																				
Sem vs Acc S1	1.606	.590	2.653	.369	3.542	.154	.118	.138	2.001	.567	.811	.889	2.801	.319	.000	.994	NA	NA	.000	.996
Sem vs Oth S1	8.074	**<.001**	8.710	**<.001**	4.553	.069	2.428	.282	4.429	.107	18.927	**.014**	144.285	**.001**	6.794	**.007**	NA	NA	64.614	.271
Model 2b – including responses from Session 2 as predictors																				
Contrasts																				
Sem vs Acc S2	.000	.994	NA	.994	49.653	**.002**	NA	.996	.839	.894	1.483	.757	5.753	.166	.573	.654	5.508	.071	.000	.169
Sem vs Oth S2	4.263	**.004**	18.653	**<.001**	171.73	**<.001**	1.624	.541	42.929	**.001**	33.944	**.010**	103.25	**<.001**	1.877	.351	3.525	.139	NA	NA